Hindawi Publishing Corporation
International Journal of Distributed Sensor Networks
Volume 2014, Article ID 461780, 9 pages
http://dx.doi.org/10.1155/2014/461780

Research Article

Investigating How User's Activities in Both Virtual and Physical World Impact Each Other Leveraging LBSN Data

Zhiwen Yu,[1] Yue Yang,[1] Xingshe Zhou,[1] Yu Zheng,[2] and Xing Xie[2]

[1] *School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China*
[2] *Microsoft Research Asia, Beijing 100080, China*

Correspondence should be addressed to Zhiwen Yu; zhiweny@gmail.com

Received 5 October 2013; Accepted 15 December 2013; Published 19 February 2014

Academic Editor: Jong Hyuk Park

In this paper, we investigate how user's online behavior (e.g., making friendships) and their offline activity (e.g., check-ins) affected each other by leveraging the data collected from LBSN. First, we use vectors to represent nodes and define popularity entropy for each node to weigh their popularities and the impact on forming a new edge. Then, we propose an algorithm to calculate the weight of each edge based on our findings that the more overlap of linked nodes they have, the heavier weight the edge has and the more popular the nodes in their overlap are, the lighter the weight of the edge is. Finally, we conduct link prediction by using the random walk with restart method considering the effect of every node and every edge. Experimental results show that user's activity in virtual world and physical world do really have great impact on each other.

1. Introduction

The advent of GPS-enabled smart phones has had a significant impact on the development of location-based services. Combined with user's growing interests in on-line social networks, it leads to the emergence of many location-based social network (LBSN) services, such as Foursquare (http://foursquare.com/), Brightkite (http://brightkite.com/), and Loopt (http://www.loopt.com/). Millions of users have already been attracted to such kind of services, with which they can leave digital footprints [1], known as check-ins, at the places where they have been to. This service also makes it possible for users to share their locations, photos, and tips with friends anytime and anywhere. Many users prefer to share their trajectories on LBSNs rather than to directly tell people where they are on traditional microblogs. The possible reason is that LBSNs provide much stronger social security, that is, user's check-ins can only be seen by a small group of people specified by the user themselves, while user's microblogs are public to anyone who has followed them on microblogging websites. When a user publishes where they are, they may worry about revealing too much personal information.

As a classical problem in complex network [2], link prediction became a hot topic again when all the nodes are replaced by humans instead of physical network nodes. A considerable amount of attention has been devoted to this problem in social networks and a lot of consensuses have been reached. The most widely accepted one is that, for any two users, the more friends they have in common, the more likely that they would become friends in the coming future. This naive idea really works and has been used in link recommendation by different on-line social networks. However, based on a sociology theory called homophily [3] which emphasizes the importance of similarity in the process of forming a new link between any two strangers, it is intuitive to infer that similar people usually go to similar places, which is to say if any two users often visit the same or similar places, they are likely to be friends. However, it has not been well explored yet because large-scale user's movements in physical world are difficult to monitor.

In this paper, we use the data crawled from Foursquare to explore how users' activities in virtual world and physical world impact each other, particularly how users' movements affect users' online friendships and vice versa. As a location-based social network, Foursquare contains two kinds of

information: one is user's social graph in visual world and the other is user's check-ins in the physical world. However, it is difficult to get check-ins directly for security reasons; thus we use tips instead, which contain not only time and location as check-ins but also some comments or messages created by users.

We use vectors with different amount of elements to represent different users and places. Each element in a vector represents an entity, which can be another user or a location linked to the user. If any two users are friends or a user has been to a place, we can say that the two users are linked to each other or the user and the place are linked to each other, which means that there is an edge between them in graph made up by users and locations. In this way, we store all the information and edges in the graph consisting of users and locations that are regarded as nodes.

All the vectors representing users and locations make up a graph, which is called global graph, while a certain user's friends and his/her visited locations make up the user's local graph. It is known to all that many ordinary users may follow many famous movie stars or famous writers in the on-line social network, but for any two of this kind of ordinary users, they are not much less possible to be friends in the future. The reason is obvious that famous users usually have very large local graphs linking them to many users, and these users do not have much in common with each other. This phenomenon also exists among the users who visited a very popular place. For example, there are numerous visitors visiting the Imperial Palace every day, but only very few users may be friends. On the contrary, users who are linked to unpopular locations and unpopular users are highly likely to be friends in the future. In order to portray this characteristic to be used in link prediction, we define a weighing entropy to weigh every node so that we can take all the information that each node contains when predicting users' relationships and movements. Finally, we conduct link prediction by using the random walk with restart method considering the effect of every node and every edge.

In the remainder of this paper, we first discuss related works in Section 2 and give several basic definitions and our motivation in Section 3. In Section 4, we analyse the statistic features of the data. The link prediction and experimental results are presented in Sections 5 and 6, respectively. Finally, we conclude the paper in Section 7.

2. Related Work

The link prediction problem in social networks has attracted a considerable attention since it was introduced by Liben-Nowell and Kleinberg [4]. Link prediction was first explored in the complex networks and when more and more users' started joining on-line social networks, researchers started to analyse users' activities in on-line social networks with different backgrounds and different purposes. In fact, we can view social networks as special complex networks whose nodes are entities (users or locations) and edges represent interaction, collaboration, or influence between entities. Before introducing the studies in link prediction, firstly we try to define this problem as follows.

Given a snapshot of a graph representing social network at time T_0, we seek to accurately predict the new edges that will be added to the graph during a certain time interval from time T_0 to a given future time T_1.

Pondering over this problem deeply, we can easily find that it contains two questions. The first one is that by what features intrinsic to the network itself the social network evolves. The second one is how the social network evolves. The key to get the answer is to find the factors that have the greatest impact on the network's evolution. We try to summarize the related work according to the factors that were used in link prediction.

2.1. Factors with User's Movements in Physical World. To predict the occurrence of new edge between any two users in a social network, the first step is to find the similarities between them. Many researchers tried to mine user similarity from users' movements. Users' movements in the physical world usually show their travelling trajectories in daily life and represent users' activities in the physical world directly. When users are at a certain place, they can only do certain things; for example, users can only watch movies in a movie theatre while have dinner at restaurants. Furthermore, similar people go to similar places with similar visiting sequences at similar time. Based on these ideas, issues focusing on mining users' similarities from users' movements have largely been explored. Li et al. [5] hire several volunteers from different countries and different cities equipped with GPS loggers to collect their trajectories for several months to mine similarities among these volunteers. They firstly detect stay points from trajectories considering the length of user's staying time and the distance among different GPS points and transferred user's raw GPS trajectories into the sequences of stay points. They build a hierarchical graph with three levels, which have different ranges of areas, to model user's location history. They argue that except for colocations, which refer to the locations that are visited by related users, among the users, visiting sequences make more sense and carry more meaningful information in mining users' similarities. Then, they tried to find similar sequences with different lengths among users and compute the similarity across multilevels. This work has made a great process in mining user similarity from physical world resources. However, it is difficult to get a large-scale nonvolunteer user's trajectories, partly because it is impossible to let everyone equip with a GPS logger and partly due to privacy issues. Therefore it is hard to apply this idea to find similarities among a large scale of users.

As more and more users use cell phones, researchers try to infer social ties from cellular network data. Cho et al. [6] use data from Brightkite (http://brightkite.com/) and Gowalla (http://gowalla.com/) along with a dataset of cell phone location trace data to understand the basic laws that govern users' motion and dynamics. They find that there basically exist two features in users' travel style. One is that users' short-ranged travel is geographically and temporally periodic and has nothing to do with social network structure; the other is that users' long-distance travel is largely affected by social network ties. More than 50% of user's movements can be

explained by periodic behaviour while less than 30% of user's movements can be explained by social ties. Thus, they build a periodic and social mobility model to predict individuals' mobility combining periodic short-ranged movements with social-ties related travels. Their model has three different components capturing the feature of user's regularly visiting spatial locations, the temporal movement between these locations, and user's movements influenced by social ties, respectively. The model has acceptable performance in predicting user's mobility. However, it only explores how user's mobility is influenced by features like social ties and periodic activities but pays no attention to how user's mobility can impact other aspects of user's life.

There are also many research works making use of users' trajectories in physical world for other interesting purposes. For instance, Lian and Xie [7] proposed a novel location naming approach to automatically provide concrete and meaningful location names to users based on their current location, time, and check-in histories.

2.2. Factors with User's Relation in the Virtual World. Nowadays, many people spent more and more time on on-line social networks such as Facebook and Twitter, keeping in touch with existing friends, getting to know new friends, and sharing ideas and resources. Many factors might have impact on the evolution of the user's social network. Xiang et al. [8] try to model relationship strength using users' profiles and interactions among users using the data from Facebook and LinkedIn. Before building a model, they first assume that people tend to form ties with people having similar characteristics and relationship strength has impact on online interactions, both on nature and frequency. They develop an unsupervised latent variable model to estimate relationship strength from interaction activities, which can be communication, tagging or something else, and user similarity extracted from users' profiles. The estimated relationship strengths result in a weighted graph of which the spurious links have been downweighted while the important ones have been highlighted. Finally, these weighted links can be used to increase the accuracy of social network mining tasks, including link prediction. Tang et al. [9] try to do link prediction using data across heterogeneous networks. The main question is how to bridge the available knowledge from different networks to help infer different types of social relationships. Their main ideas come from several social psychological theories such as social balance, structural hole, and social status. They proposed a transfer-based factor graph model incorporating social theories into a semisupervised learning framework used to transfer supervised information from the source network to help infer social ties in the target network. As we all know almost every user has accounts on different online social networks and different networks contain different aspects of information of a user. This work represents a new and interesting research direction in making full use of user's online activities.

Besides the data from online social network, emails can be regarded as users' activities in the virtual world. Researchers from Google [10] made use of users' implicit social graph, which is formed by users' interactions with contact and groups of contacts in Gmail, to do friendship recommendation. They also proposed an interaction-based metric for estimating a user's affinity to their contacts and groups. Their experiments showed that both implicit social graph and interaction based affinity were important in suggesting friends.

2.3. Factors with Mixed Resources. As online social networks and user's movements contain user's activities and relationships in virtual and physical world, respectively, if we try to understand user's activities from only one point of view, we can only get biased results. To get better and unbiased understanding of user's activities in both circumstances, several researchers turn to combine information from these two aspects to do link prediction. Cranshaw et al. [11] first analyze the social context of a geographic region from a set of location-based features including location entropy. Then, they compose a model to predict the friendship between any two users by analyzing their location trails. Finally, they show a positive relationship between the colocation histories and the social ties that the user has in the network. Their work proves that offline mobility has impact on user's online activities. Wang et al. [12] try to find the relation between human mobility, social ties, and link prediction. Mobile communication records are regarded as the representation of social ties while user's moving trajectories are extracted from cellular network. The authors try to predict the social ties with different datasets, each of which contains different proportion of user's mobility.

Scellato et al. [13] described a supervised learning framework which exploits prediction features which they extracted from data of Gowalla to predict new links among friends-of-friends and place-friends and showed that the inclusion of information about places and related user activity offers high link prediction performance.

User's friendship usually can be crawled from online social networks, but user's mobility is very hard to get. Previous works usually tend to use cellular network data or hire volunteers to collect trajectories. However, cellular network data is often of low accuracy, while hiring volunteers to collect data is only suitable for a small scale of study. No matter which one of these two resources was used, researchers have to spend much time in extracting locations with semantic or geographic context from the raw data. In fact, we care much more about where the user has been to instead of how he/she got there. Location-based online social networks offer such kind of information as users often tag the places they have visited. In this way, we can do link prediction and study the impact that user's mobility and social ties have on each other from a new point of view.

3. Preliminary

In this section, we first give several definitions and then introduce our main motivations. To simplify the explanation, we view locations and users as the same which is determined by our motivations.

3.1. Definitions

Definition 1 (node). A node v refers to a user or a location that the network contains. Users and locations have the same status in this paper. Each node v is associated with a unique ID, $v \cdot id$, which starts from 1 and ends at the number of total nodes.

Definition 2 (edge). If two users, represented by node j and node k in the network, are friends, there will be an edge $e_{j,k}$ between them, showing that they are connected to each other. Similarly, if a user has visited a place, there will also be an edge between them. Each edge $e_{j,k}$ is undirected and associated with two terminal points, $e_{j,k} \cdot v_j$ and $e_{j,k} \cdot v_k$, referring to the two nodes (node j and node k) it connects. $e_{j,k} \cdot w$ is the weight of the edge determined by the popularity entropy of two nodes. Furthermore, a user can visit a place for several times but only can be friends with another user for only once, which means that part of edges may occur more than once. To record this feature, the last attribute of each edge $e_{j,k}$ is $e_{j,k} \cdot t$ recording the times the edge occurred. For edges connecting users, the value is 1, but for those connecting users and locations, the value may be large than 1.

Definition 3 (global social graph). A global social graph $G(V, E)$ contains all the information and represents the structure of a whole location-based social network containing two types of elements, node v and edge e. It is an undirected graph.

Definition 4 (local social graph). Node l's local social graph $G_l(V_l, E_l)$ may contain the complete friendship and all the visited places of a user, or all the users who have been to a particular place. All nodes in the local social graph G_l are connected to the node l directly; that is, only one-hop connections exist in local graph. The graph is also an undirected graph.

Definition 5 (cofriend). For any two friends, A and B, of a user, the user is a cofriend cf of them.

Definition 6 (colocation). If two users have been to the same place, the place is their colocation cl. We do not require that the two users must visit the place at the same time.

Definition 7 (popularity entropy). Popularity entropy pe is used to weigh a node's popularity among all the nodes and has impact on the strength of links connected to the nodes.

3.2. Motivations. Our motivation is quite simple and direct. Most existing research works focus on predicting the link among friends' friends or between friends and friends' visited locations, all of which are nodes that are no more than 2 hops away from each other. However, we argue that new links within 2 hops away are just a small part of all the newly added edges and there are a large amount of newly added edges occurring between the nodes whose distance is more than 2 hops. Figure 1 shows the proportion of the two kinds of new added links occurring during the period from September 2011 to February 2012 in our dataset collected from Foursquare.

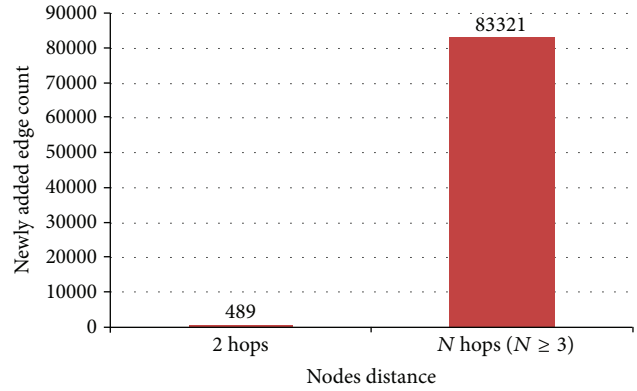

Figure 1: Distance of nodes that newly added edges link.

We can see from Figure 1 that very few new links are added among the nodes within 2-hop distance while most new links are added among the nodes which are farther away from each other. The result shows that if we only pay attention to the new links within 2 hops, much information will be lost and high precision of link prediction cannot be achieved.

It is well accepted that for any two unconnected users, the more friends they share, the more likely that they would be friends in the near future. But we may have a common sense that famous users, such as movie stars like Justin Bieber, or some dignitaries, for example, Obama, usually have a very large social network in those online social networks. If any two users whose most friends are the famous users on the network, do they have large probability to be friends in the future? It is probably not. We argue that except the amount of common friends any two users have, the scale of common friends' social network also has impact on forming a new link between two strangers.

The last motivation is that since online social network has become a part of life, we want to analyse the interaction between our daily life and the online social network. To explain it in detail, we divide user's activities into two parts: the first part is user's activities in the offline world, for example, moving to a new city or meeting with new friends and so forth, while the other part is user's online activities, such as adding new friends or commenting on friends' photos. What kind of influence these two kinds of activities have on each other is largely unexplored.

Before further introducing our work, we introduce two basic hypotheses in this paper as follows:

Hypothesis 1. The larger the scale of overlap that any two users' relationship has in virtual world is, the more likely that they will have larger scale of overlap in check-ins in physical world.

Hypothesis 2. The larger the scale of overlap that any two users' check-ins in physical world have is, the more likely that they will have large scale of overlap of relationship in virtual world.

Table 1: Amount of newly added edges in different groups.

Group	Amount of shared friends	Amount of shared friends' friends	Amount of new added edges
1	10.33	25.74	1.41
2	11.56	14.79	3.75

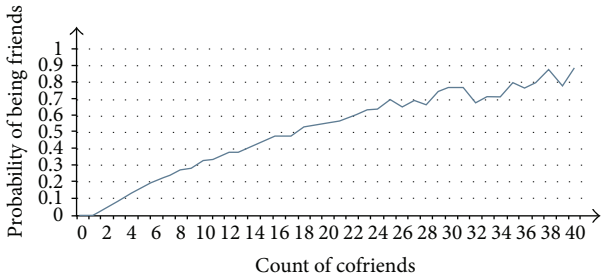

Figure 2: The relationship between probability of being friends and the amount of cofriends.

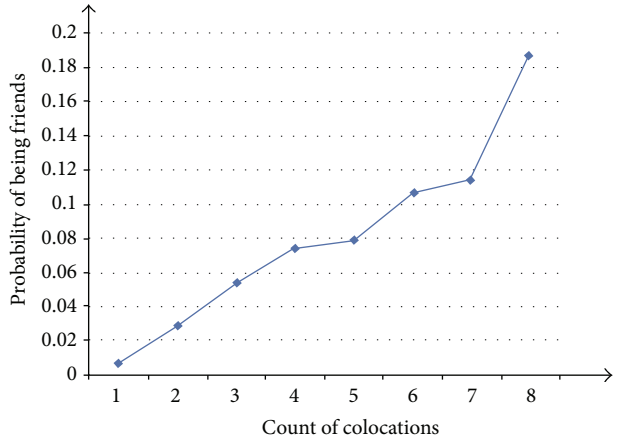

Figure 3: The relationship between probability of being friends and the amount of colocations.

These two hypotheses come from the goal of exploring the impact of user's activities in physical and virtual world on each other. Our further experimental results will show that these two hypotheses are reasonable and meet the phenomenon in reality.

4. Observations

We use the dataset crawled from Foursquare. The dataset contains 31524 users, 51265 venues (locations), and users' tips about these venues from July 2011 to May 2012. In this paper, we take user's check-ins extracted from tips (time and geographic information) as user's activity in the physical world and the evolution of user's friendship as user's activity in the virtual world. We here present the statistic characteristic of the data collected from July 2011 to February 2012.

Firstly, we analyze how the amount of shared friends has influence on users' social ties. Figure 2 shows the relationship between the amount of shared friends that any two users have and their probability of being friends.

In Figure 2, we only show the records of samples with a relatively larger scale. It is easy to find that the probability of being friends almost rises as the amount of shared friends grows. However, it is also obvious that the curve shakes with the increment and the probability does not grow purely. How does it happen? Dose the probability not grow as the amount of common friends rises? Just as our motivation, it is not always true. To verify the idea, we choose two groups of users. In Group 1, the shared friends of any two users are almost the users who own a relatively larger social network, while in Group 2, shared friends are with a relatively smaller social network. We analyse the average amount of new links added among these users from September 2011 to February 2012 in different groups and the results are definitely different, as shown in Table 1.

It is not surprising to find that there are only very few new links added in the first group but much more in the second group. Just as mentioned in our motivation, users with a large amount of friends are usually some kind of famous people in the social network and anyone who is interested in them can add them as friends. As a result, the links between the famous user and their friends are quite weak and cannot be used as the support to predict new links. But for the less popular users, since their social networks are relatively smaller, their friends are closely connected with them and the links are the strong supports for predicting new links.

Since we take user's check-ins as users' activities in physical world and evolution of their friendships as activities in virtual world, and our goal is to find whether user's activities in virtual and physical world have impact on each other, we have to investigate two topics. The first one is that if two users have N colocations, we want to know what the probability of the two users being friends is.

From Figure 3, we can observe that as the number of colocations of two users rises, the probability of being friends increases obviously. However, we can observe some shakes clearly again. The reason why the shakes occur is similar to that in Figure 2. There are always some places attracting numerous people while some others are visited by a very small group of people. The popularity of a place has great impact on the strength of the links between the place and the attracted users. The more the people visit a place, the more popular the place is, the weaker the links between the place and user are and vice versa. As the number of colocations rises, the proportion of popular colocations changes and the amount of pairs of users who have colocations changes, which is the reason why the shakes occur.

The second topic we want to investigate is that, as the number of cofriends increases, whether the probability of having colocations will rise. The result is shown in Figure 4.

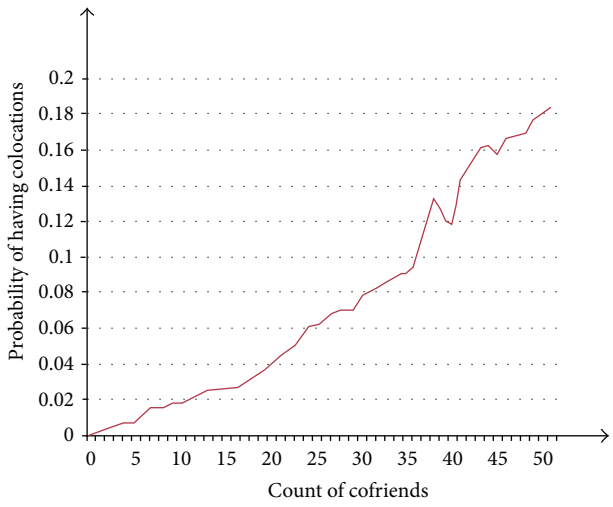

Figure 4: The relationship between probability of having colocations and the count of cofriends.

We can observe that when the number of cofriends increases, the probability of having colocations also increases, but there are also some shakes. The explanation is similar to the above one.

All the static analysis results show that it is not proper to consider the amount of colocations or cofriends only when predicting new links. Each member of colocations and cofriends contains much information about the links and helps us to understand the existing links better. But how to translate the information into a form that can be used in link prediction remains a problem. In the next section, we will try to solve it.

5. Impact Detection

We try to study how user's friendship and their check-ins impact each other. Our main task is to predict the new edges (links) in the graph with nodes representing users and locations. It is well accepted that the more information we obtain and use when we try to predict new links, the higher precision we can get. However, due to the limits of the space of storage and time, it is not practical to store all information. It is not a good way to simply count the number of colocations and cofriends, since each element of them also contains some factors that have impact on the addition of new links. In this paper, we propose an effective way to handle the issue.

Every node in the global graph is connected with many nodes and these nodes may be connected with many other nodes. Each node's popularity is determined by its connection with the rest of nodes. We use a vector to represent each node. There are a very large number of nodes including users and locations in the global graph but only a few nodes are connected with each other, which means that the global graph is very sparse. If we store all the edges in the graph without considering whether they really exist, the most elements in every vector will be 0 meaning that the nodes are not connected and most information in the vectors is meaningless. In fact, only the connected nodes and their edges make sense in link prediction. Since each edge stores the information about its weight, its occurrence times, and the two nodes it connects, a node's vector which contains all the edges connected to the node can store all the information that the node has. By doing so, the topology of the social networks is stored with high coefficient of storage space utilization and it also ensures that all the nodes in the social network have the same status.

If the social network contains N nodes in total, for node i connected with n edges, its vector, v_i, is defined as follows:

$$v_i = v\left(e_{i,j_1}, e_{i,j_2}, \ldots, e_{i,j_n}\right), \quad i \in [1, N], \ j_m \neq i, \ m \in [1, n], \tag{1}$$

where $e_{i,j}$ $(i \neq j)$ represents the edge existing between node i and node j and $e_{i,j}$ is the same with $e_{j,i}$ $(i \neq j)$, because we treat the graph of social network as on undirected graph in this paper. All the elements' initial values of weight are set to be 1 and the final value will be larger than 0 in every vector. Since all the weights are associated with the nodes' popularity entropy, we introduce a method to calculate entropy as shown in Algorithm 1. $e_j \cdot t$ means that the number of the edges of e_j appears in the dataset.

For a node i, we first get the sum count n of its vector's elements, then we calculate the sum value of occurrence times of the edges linked to node i. According to our hypothesis, for any node, its popularity is proportional to the amount n of other nodes it is connected with and the occurrence times of edges that are linked to the node. In order to get normalized weight for each node, the popularity entropy pe_i for node i is computed as follows:

$$\mathrm{pe}_i = n * \frac{\mathrm{value}}{\mathrm{Sum}_E}, \tag{2}$$

where the sum Sum_E of edges is used to complete the normalization. To explain the algorithm in detail, the basic idea of calculating the weight of each edge is summarized as follows.

For two nodes that have been connected to each other in the location-based social network,

(1) the more shared nodes they have, the stronger the tie between them is; that is why we use the sum of the values got from each element as edge's weight between the two nodes;

(2) the more less-popular nodes they share, the stronger the tie between them is; that is why we use the reciprocal of the nodes' popularity entropy as a factor that has impact on the strength of tie between the two nodes;

(3) the more times an edge occurs, the more important the linking nodes are for each other; so we use the sum of the occurrence times they connected with their shared nodes to weigh the importance of the shared nodes for the two nodes.

As the values of $e_{j,i} \cdot w$ and $e_{i,j} \cdot w$ are the same, we can compute them at the same time. To keep the computation in

Input: a collection of nodes' initial vectors V, the sum of edges Sum_E
Output: a collection of nodes' popularity entropy PE = {pe}
(1) $i \leftarrow 1, N \leftarrow |V|$
(2) While ($i < N + 1$)
(3) $\quad \text{pe}_i \leftarrow 0, n \leftarrow \|v_i\|, value \leftarrow 0$
(4) $\quad$ foreach e_j in v_i
(5) $\quad\quad value = value + e_j.t$
(6) $\quad\quad j \leftarrow j + 1$
(7) $\quad \text{pe}_i = n * value/\text{Sum}_E$
(8) $\quad i \leftarrow i + 1$
(9) $\quad$ return pe_{i-1}

Algorithm 1: Popularity entropy calculation.

Input: a collection of nodes' vectors V, a collection of nodes' popularity PE, the sum of edges Sum_E
Output: a collection of nodes' vectors $V = \{v\}$,
(1) $i \leftarrow 1, N \leftarrow |V_{\text{ini}}|, Initial(V)$
(2) While ($i < N + 1$)
(3) $\quad$ foreach $e_{i,j}$ in v_i
(4) $\quad\quad$ if $e_{i,j}.vj > i$
(5) $\quad\quad\quad j \leftarrow e_{i,j}.vj, P \leftarrow CommonNodes(v_t, v_j), e_{i,j}.w = e_{j,i}.w = 0$
(6) $\quad\quad\quad$ foreach p in P
(7) $\quad\quad\quad\quad m \leftarrow \|v_p\|, n \leftarrow \text{sum}(e_{i,p}.t, e_{j,p}.t)$
(8) $\quad\quad\quad\quad e_{j,i}.w = e_{i,j}.w = e_{i,j}.w + reverse(\text{pe}_p) * n/m$
(9) $\quad i \leftarrow i + 1$
(10) $\quad$ return v_{i-1}

Algorithm 2: Edge final weight calculation.

a controllable style and save both the time and space, we let each node only compute the strengths of ties with the other nodes having larger node ID than it does. For any node i, we check the elements in its vector one by one. For an element $e_{j,i}$ in node i's vector, we first check the id $v_j \cdot id$ of the other node that node i connects; if $v_j \cdot id$ is larger than $v_i \cdot id$, then we store the node's id as j and try to get the shared nodes of node i and node j. Then, for every shared node p, we calculate its vector's count m and the sum of $e_{j,p} \cdot t$. According to our hypothesis the occurrence times of edges linking to shared nodes are proportional to the weight of edge $e_{i,j}$, while the popularity entropy of shared nodes is inversely proportional to the weight of that edge. So each shared node's contribution to the edge $e_{i,j}$ is computed by the following formula:

$$C_{p-ij} = \text{reverse}\left(\text{pe}_p\right) * \frac{n}{m}. \tag{3}$$

C_{p-ij} represents the shared node p's contribution to the edge $e_{i,j}$, and reverse (pe_p) represents the reverse value of pe_p. By computing the sum of all shared nodes' contribution, we get the value of edge $e_{i,j}$. All these processes are shown in Algorithm 2. The algorithm is different from TF-IDF [14]. Using TF-IDF algorithm, we should firstly specify one kind of nodes as a document containing many words and the others are chosen to be words. However users and locations have the same status in the social network; choosing any one of them as document will destroy the balanced status.

Using Algorithm 2, we can get the weights of all the edges existing in the social network. After getting the popularity of each node and weight of each edge, we use random walk with restart [15] as the model to do link prediction.

6. Experimental Results

We divide the crawled dataset into two parts: the first part starts from July 2011 to February 2012 and the second part start from February 2012 to May 2012. We use the first part data as training set and the second part as test data. We would like to use the test set to verify the effectiveness of our method in improving the accuracy of link prediction. To go one step further, we also want to show that user's activities in virtual world and physical world have impact on each other, which can be shown in the result of link prediction.

First, we would like to experimentally evaluate the capability of our method to compute the strength of ties between real entities, for example, between user and user or between user and location. We use Algorithm 2 to calculate the weights of existing edges and the possible weights of edges between nodes that are not really connected to each other. In other words, we suppose that any two nodes in the social network were connected to each other and we

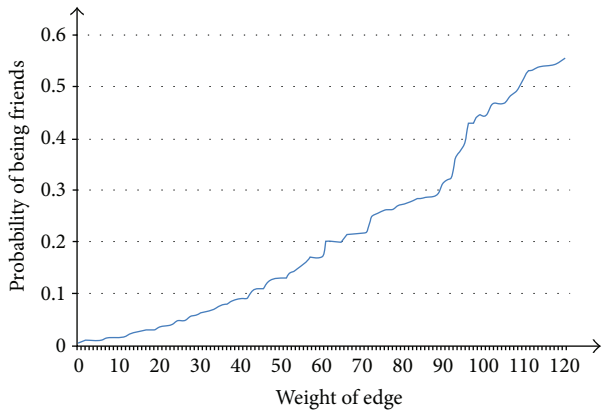

Figure 5: The relationship between probability of being friend-nodes and the weight of edges.

Table 2: Contents of different graphs.

Graph Feature	AW	UW	UUW	ULW
User' relation	*	*	*	—
User's tips	*	*	—	*
Weighted or not	*	—	*	*

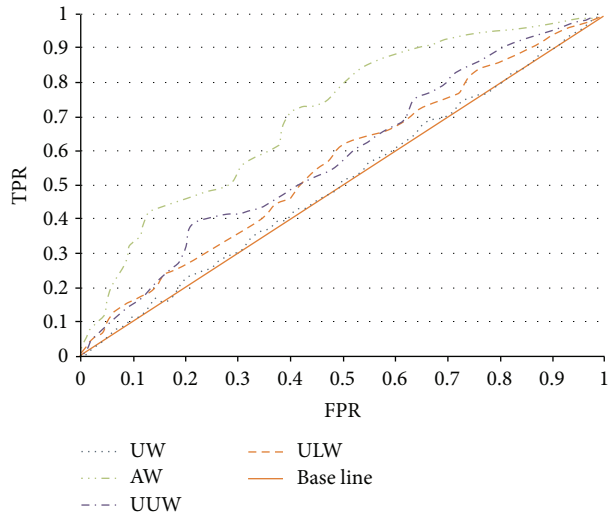

Figure 6: Link prediction results using different information (the graph with all weighted edges (AW), the graph with all edges' weights to be 1 (UW), graph with only user-user edges (UUW), and graph with only user-location edges (ULW)).

calculate the weights of all edges even though some of them may not exist. Then, we view the pair of nodes that are really connected as friend-nodes and finally get the relationship between probability of being friend-nodes and the weight of edges. The results are shown in Figure 5 where the whole dataset is used (i.e., July 2011 to May 2012).

From Figure 5, we can observe that for any two nodes, the probability that they are connected with each other varies directly with the weight of the edge linking them. Furthermore, there is almost no shake before the ending part of the curve. Comparing to Figure 2, this result verifies our idea that nodes' relations are not only impacted by the amounts of shared nodes but also the amounts of connected nodes each shared node owns. Since the probability varies along with the weight of edge, the curve also shows that it is really necessary to take the existing edges' weights into account when performing link prediction instead of counting the amount of shared nodes only.

To analyze the impact that the physical world and virtual world have on each other, we conduct the experiments on different graphs, for example, the graph with all weighted edges (AW), the graph with all edges' weights to be 1 (UW), graph with only user-user edges (UUW), and graph with only user-location edges (ULW). Details are as shown in Table 2, where "*" means that the corresponding vertical line contains the information in the related horizontal line while "—" means that is does not. For instance, UW contains the information of user relation and tips (i.e., check-ins).

We use the four graphs to show how different types of information have impact on the result of link prediction. Comparing the results of AW and UW, we intend to see how the weighted information can be used to improve the performance of link prediction. The purpose of comparing the results of AW and UUW is to find out how a user's activities in the physical world can influence his/her activities in the virtual world, while comparing the results of AW and ULW is just for the inverse purpose. We use the graph in MAY 2012 as ground truth. The new links are predicted based on the previous observations and the graph in February 2012. Finally, we use ROC curve to demonstrate each graph's result as shown in Figure 6.

From Figure 6 we can observe that the performance of the unweight graph (UW) is the worst while the weighted graph with all kinds of nodes and edges (AW) performs the best, which means that our method of calculating each edge's weight is basically effective. Furthermore, UUW containing the weighted user-user edge only performs better than UW but much worse than AW, which shows that using both the information in physical world and virtual world in link prediction can provide much better results. This also proves that user's activities in physical world have impact on his/her virtual world activities. On the other hand, we can draw a conclusion that user's activities in the virtual world can influence user's activities in the physical world by comparing the performance of AW and ULW (the graph with only weighted user-location edges).

7. Conclusion

In this paper, we present a study investigating how user's activities in virtual world and physical world impact each other. We use information vector to represent nodes and to store all the edges that are linked to every node. We define popularity entropy to weigh each node's popularity according to the amount of nodes it connects and the occurrence times of edges linked to it. Then, we calculate each edge's weight considering its linking nodes' popularity entropy and their shared nodes' features. In this way, we get a weighted graph

and use it for link prediction. The results show that user's activities in virtual world and physical world do have impact on each other and using both virtual world and physical world information can improve the accuracy of link prediction.

In the future, we plan to detect and analyze communities formed by the nodes of online social network to see whether user's activities in virtual and physical world have impact on community dynamics. Real applications that employ the results will also be a future work of this study.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was partially supported by the National Basic Research Program of China (no. 2012CB316400), the National Natural Science Foundation of China (nos. 61103063, 61222209, 61373119, and 61332005), the Specialized Research Fund for the Doctoral Program of Higher Education (no. 20126102110043), and Microsoft Corporation.

References

[1] D. Zhang, B. Guo, and Z. Yu, "The emergence of social and community intelligence," *IEEE Computer*, vol. 44, no. 7, pp. 21–28, 2011.

[2] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[3] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: homophily in social networks," *Annual Review of Sociology*, vol. 27, pp. 415–444, 2001.

[4] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[5] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma, "Mining user similarity based on location history," in *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '08)*, pp. 298–307, ACM Press, November 2008.

[6] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*, pp. 1082–1090, ACM Press, August 2011.

[7] D. Lian and X. Xie, "Learning location naming from user check-in histories," in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '11)*, pp. 112–121, ACM Press, November 2011.

[8] R. Xiang, J. Neville, and M. Rogati, "Modeling relationship strength in online social networks," in *Proceedings of the 19th International World Wide Web Conference (WWW '10)*, pp. 981–990, ACM Press, April 2010.

[9] J. Tang, T. Lou, and J. Kleinberg, "Inferring social ties across heterogeneous networks," in *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM '12)*, pp. 743–752, ACM Press, February 2012.

[10] M. Roth, D. Deutscher, G. Flysher, I. Horn, and A. Leichtberg, "Friends using the implicit social graph," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)*, pp. 233–242, ACM Press, July 2011.

[11] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh, "Bridging the gap between physical location and online social networks," in *Proceedings of the 12th International Conference on Ubiquitous Computing (UbiComp '10)*, pp. 119–128, ACM Press, October 2010.

[12] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabási, "Human mobility, social ties, and link prediction," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*, pp. 1100–1108, ACM Press, August 2011.

[13] S. Scellato, A. Noulas, and C. Mascolo, "Exploiting place features in link prediction on location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*, pp. 1046–1054, ACM Press, August 2011.

[14] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management*, vol. 24, no. 5, pp. 513–523, 1988.

[15] H. Tong, C. Faloutsos, and J.-Y. Pan, "Fast random walk with restart and its applications," in *Proceedings of the 6th International Conference on Data Mining (ICDM '06)*, pp. 613–622, IEEE Press, December 2006.